LOGISTIC AND MULTIPLE REGRESSION: THE TWO-STEP APPROACH TO ESTIMATING COST GROWTH

THESIS

Daniel C. Genest, First Lieutenant, USAF

AFIT/GCA/ENC/04-01

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U. S. Government.

AFIT/GCA/ENC/04-01

LOGISTIC AND MULTIPLE REGRESSION: THE TWO-STEP APPROACH TO ESTIMATING COST GROWTH

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cost Analysis

Daniel C. Genest, BS

First Lieutenant, USAF

March 2004

AFIT/GCA/ENC/04-01

LOGISTIC AND MULTIPLE REGRESSION: THE TWO-STEP APPROACH TO ESTIMATING COST GROWTH

Daniel C. Genest, BS
First Lieutenant, USAF

Approved:

/Signed/	30 Jan 04
Edward D. White, PhD (Chairman)	date
/Signed/	30 Jan 04
Michael A. Greiner, Major, USAF (Member)	date
/Signed/	30 Jan 04
Michael J. Seibel (Member)	date

Abstract

This study seeks to predict cost growth in major DoD acquisition programs using logistic and multiple regression. Specifically, this research uses logistic regression to determine whether or not cost growth will occur in a program and if so, then uses multiple regression to determine to what extent that cost growth will occur. We compile data from all defense departments using the Selected Acquisition Reports presented between 1990 and 2002. We combine the efforts of previous research and focus our study on cost growth in research and development dollars for the Engineering Manufacturing Development phase of acquisition. For the logistic regression portion of our research, we produce a seven-variable model that accurately predicts 72 percent of our randomly selected validation data. For multiple regression, we produce a six-variable model that accurately predicts the amount of cost growth incurred for 91 percent of those programs that do incur cost growth. We conclude that the two-step regression methodology offers a significant advantage over traditional methods by removing those data points that do not incur cost growth. We further conclude that there is no significant advantage gained by either isolating each cost variance category individually or by combining these categories.

Acknowledgements

I would like to thank Dr. Edward White for his endless guidance and patience throughout this venture. I would also like to thank Major Michael Greiner and Mr. Michael Seibel for their time and energy in supporting my research effort and for providing invaluable feedback. Next, I would like to thank Captain Vince Sipple for providing such a solid foundation on which we base this research and Captain John Bielecki for his help in getting up to speed during the first few months. Also, I would like to thank the U.S. Air Force for providing me with the incredible opportunity to be a student at AFIT and obtaining a master's degree in cost analysis.

Last, but certainly not least, I would like to thank my wife for her encouragement, support and patience. I would also like to thank my wife and son for providing a constant source of motivation and perspective.

Daniel C. Genest

Table of Contents

List of Figures

Page

List of Tables

Page

LOGISTIC AND MULTIPLE REGRESSION: THE TWO-STEP APPROACH TO ESTIMATING COST GROWTH

I. Introduction

General Issue

Cost growth in major weapon systems procurement has been a problem that has plagued the Department of Defense (DoD) ever since the data necessary to compute cost growth has been recorded. Since the formalization of the Selected Acquisition Report (SAR) by Congress in 1969, cost growth within DoD has averaged roughly 20 percent (Drezner, 1993: xiii). The persistent inability to reduce cost growth through more accurate cost estimates creates a serious problem as Congress uses these estimates to allocate resources provided in good faith by the American public. Congress faces the task of providing for national defense through this allocation of resources and this responsibility often entails choosing between competing weapon systems. Naturally, the estimated cost associated with each competing system plays a significant role in the decision-making process. When these cost estimates are inaccurate, the comparisons between competing systems become inaccurate as well and subsequent decisions made by Congress will potentially be influenced in an adverse manner.

Congress has taken steps to try to control cost growth by enacting legislation, such as the Nunn-McCurdy Act, with the intention of bringing more visibility and scrutiny to programs that incur large cost increases. While the overall effectiveness of such legislation can be argued, the message to program managers is clear; Congress takes

cost growth very seriously and programs need to control their cost growth or suffer the consequences.

Cost growth occurs due to many factors that are beyond the cost estimators' ability to estimate with precision. Every program contains a certain amount of risk and uncertainty that cannot be measured in dollars yet must somehow be accounted for in the cost estimate. The cost estimator must base the estimate not only on the actual costs of the program but also on the perceived amount of risk and associated uncertainty. The ability of DoD to control cost growth depends on the ability of the cost community to assign dollar values to these risks. By accurately converting these risks and uncertainties to dollar values, the cost estimator adjusts the estimate accordingly and minimizes the effect of cost growth.

Specific Issue

Cost estimators have various methodologies available for building cost estimates and the type of methodology chosen usually depends on the type of program being estimated. In the absence of historical data, cost estimators must often depend on subjective techniques such as expert opinion to assign dollar values to risk. Whenever possible, however, estimators usually select a more objective approach, such as collecting historical data on analogous systems and formulating an estimate on the new system based on this historical data. This methodology presents a possible problem when the new program contains characteristics unique to that program only and the historical averages do not account for these differences.

As a program progresses through its life cycle, estimators update the estimates using actual costs that have been realized up to that point. These estimates tend to be more accurate since information about many of the original unknowns regarding risk and uncertainty become known. However, analysts typically measure cost growth using the original estimates formulated early in the program's life cycle. Accordingly, in order to reduce future cost growth, a more reliable, less subjective method of accounting for these risks and uncertainties must be used.

In recent years, the use of statistical regression has proven to be successful in predicting the relationships associated with cost growth. This research follows on to the work of Sipple (2002) and Bielecki (2003) and further explores the possibilities of using statistical regression to accurately estimate the dollar value associated with risk and uncertainty early in a program's life cycle. In doing so, we intend to reduce cost growth by increasing the accuracy of the original cost estimates subsequently used to compute cost growth.

Scope and Limitations of the Study

As mentioned earlier, Congress has imposed SAR reporting on major DoD acquisition programs since 1969. As such, the SARs offer the most detailed, consistent information about such programs and therefore provide the database of choice when analyzing cost growth for major DoD acquisition programs. SARs contain information necessary to identify the three cost estimates, planning, development, and current, which prove valuable in analyzing program cost growth (Calcutt, 1993:3). Also, since this

research follows and supports previous research conducted by Sipple and Bielecki, the SAR database must be used to maintain consistency among the results.

There are many different ways to measure cost growth. The two most common techniques measure cost growth as the deviation from the planning estimate (PE) to the current estimate and from the development estimate (DE) to the current estimate. These different techniques often produce vastly different results and the chosen technique usually depends on the intended use of the resulting information. Since this research focuses on the factors that cause cost growth, we define cost growth as the difference between the development estimate and the most recent current estimate available. This research analyzes programs during the Engineering and Manufacturing Development (EMD) phase in the Research and Development, Test and Evaluation (RDT&E) phase of acquisition.

The SARs separate program cost variance into seven categories: Economic, Quantity, Estimating, Engineering, Schedule, Support, and Other (Calcutt, 1993; 4). Analysts calculate cost growth that occurs after the program's baseline and attributes the growth to one of these seven categories. Since this research specifically attempts to combine the efforts of Sipple and Bielecki, it focuses only on cost growth associated with Estimating, Engineering, Schedule, Support, and Other. We do not consider Economic and Quantity cost variances since these categories, by convention, usually extend beyond the control of the cost estimator (Bielecki, 2003; 4). To ensure consistency among the results, this research also adheres to the guidelines previously set forth by Sipple and subsequently adhered to by Bielecki. These guidelines allow that "only one SAR per program is used, the most recent available, and in some instances, the most recent

available DE-based SAR is the last SAR of the EMD phase" (Sipple, 2002; 4). Further, this research limits the scope of the data to only unclassified programs since classified programs tend to contain undisclosed amounts of money within their estimate figures. Lastly, this research builds on the database used by the aforementioned researchers by updating the most recent current estimates available and including any programs that have been added to the SARs since the previous studies were conducted.

Research Objectives

As mentioned earlier, this study builds on research conducted by Sipple and Bielecki. As a result, the objectives of this research mirror those outlined in their research with only minor changes in scope. First, we use logistic regression to determine the predictive capability of certain program characteristics to forecast cost growth in the RDT&E budget during the EMD phase of development. We use logistic regression to produce a binary response. For our research, the regression produces either "Yes, the program will experience cost growth," or "No, the program will not experience cost growth." Next, we use multiple regression to determine the degree to which cost growth will occur based on certain program characteristics. Consequently, this research seeks to explore these predictive relationships so that we develop a predictive model that cost estimators may use early in a programs acquisition life cycle to ascertain potential cost growth in the EMD phase of program development (Sipple, 2002; 5).

Chapter Summary

Building on the research previously conducted by Sipple and Bielecki, this research develops a predictive model that the cost estimating community can employ

early in a program's life cycle to accurately account for the dollar values associated with that program's risk and associated uncertainty. If successful, this model will provide for more accurate development estimates and will therefore reduce the effect of cost growth in these programs.

In order to develop this model, this research uses historical data provided from the SAR database. While the SAR has some limitations, it clearly provides the most detailed and consistent source of information available for such research. This research uses logistic regression to determine whether or not cost growth will occur in a program and if so, then uses multiple regression to determine to what extent that cost growth will occur. For this research, we define cost growth as the deviation between the development estimate and the most current estimate available. Finally, this research aims to reduce overall cost growth by providing program managers and more specifically, cost estimators with a predictive model that will enable them to assess risk and uncertainty associated with a program and incorporate accurate dollar values for these risks into early cost estimates.

II. Literature Review

Chapter Overview

This chapter provides a synopsis of previous research that deals directly with historical cost growth within the DoD. We first begin with an overview of the cost-estimating process then follow with a summary of past research dealing with cost growth within the DoD. This literature review emphasizes the study by Sipple (2002) because this work contains an exhaustive review of all prior research pertaining to cost growth within the DoD. Additionally, Sipple (2002) establishes the framework for our study and the database from which we conduct our research.

Cost Risk vs. Cost Growth

This research aims to provide a predictive tool to help reduce the overall impact of cost growth within DoD acquisition programs by accurately accounting for and assigning dollar values to the cost risk associated with each program. As such, we find it important to clarify the difference between cost risk and cost growth as they relate to this study.

Within the cost estimating environment and for the purpose of this study, we define cost growth as the total cost of a system minus the originally estimated cost of that system. We define cost risk as the dollar value held in reserve to cover that predicted cost growth. Stated simply, cost risk represents the projected increase associated with uncertainties in a program while cost growth represents the actual increase incurred throughout the life of a program (Coleman, 2000: 3).

Risk Analysis

The Air Force Materiel Command's (AFMC) *Financial Management Handbook* offers the following guidance regarding risk analysis:

> Cost estimating deals with uncertainty. What the analyst attempts to do is describe in the best terms possible the probability distribution of the cost event in the future. One value for the cost estimate is the result of one prediction of that future event. Risk analysis is a careful consideration of the areas of uncertainty associated with future events. The preferred common denominator for translating risk identified in the program is dollars. *(AFMC Financial Management Handbook,* 2001:11-12)

Therefore, the cost estimator bears the responsibility of evaluating the probability distributions associated with the likelihood of each future cost event occurring, and then quantifying that likelihood into dollar values. The AFMC *Financial Management Handbook* breaks risk down into three areas: technical risk, schedule risk, and cost risk. The cost estimator assesses the probability distributions for each of these three areas and formulates a cost estimate based on that assessment. At the end of the day, however, the program manager has the final authority on which dollar values are incorporated into the final estimate sent forward (Sipple, 2002: 14).

The following three methods for developing a probability distribution are discussed in the AFMC handbook: a posteriori, a priori, and subjective judgment.

1) The first method, a posteriori, or "after the fact" relationship to past events (direct knowledge), is based on some previous occurrence such as the cost outcome of previous projects conducted by the organization. If enough samples from the past history (the population) are drawn, the probability of the next event occurring in a particular way may be estimated. A methodology like Monte Carlo simulation may also be used. The Monte Carlo simulation is conducted where the analyst determines the probability of future events by using an experimental model to approximate expected actual conditions. Such a model is fashioned from previous histories of similar projects.

2) Sometimes a distribution of possible outcomes for an event is not based on experience or sampling but on a priori, or "before the fact" theoretical probability distribution. The use of the closeness of the assumptions used in developing the theoretical distribution is to the real world situation being analyzed.
3) Many times an analyst will have to use a subjective judgment (indirect knowledge) in estimating probability. This approach relies on the experience and judgment of one or more people to create the estimated probability distribution. The result is known as a subjective probability. A distribution estimate is an analysis by one or more informed persons of the relative likelihood of particular outcomes of an event occurring. Distribution estimates are subjective. An example of this approach is the Delphi method. (*AFMC Financial Management Handbook*, 2001: 11-12)

Estimating Methods

Cost estimators have various methods at their disposal for assessing and assigning dollar values to risk. The decision as to which method to employ usually depends on how much time the estimator has to formulate the estimate. The level of detail involved in each estimate varies and depends on the purpose of the estimate and the time constraint placed upon the estimator. The estimator must then strike a balance between the level of accuracy and detail included in the estimate and the amount of time available to produce a final dollar figure. The more time the estimator has to prepare an estimate, the greater the amount of detail that can be accounted for in the estimate.

Figure 1 shows five different risk assessment techniques applied by the Ballistic Missile Defense Organization (BMDO) cost estimating community (Coleman, 2000:4). As the degree of precision required increases, the degree of difficulty as well as the time necessary to complete the estimate increases accordingly.

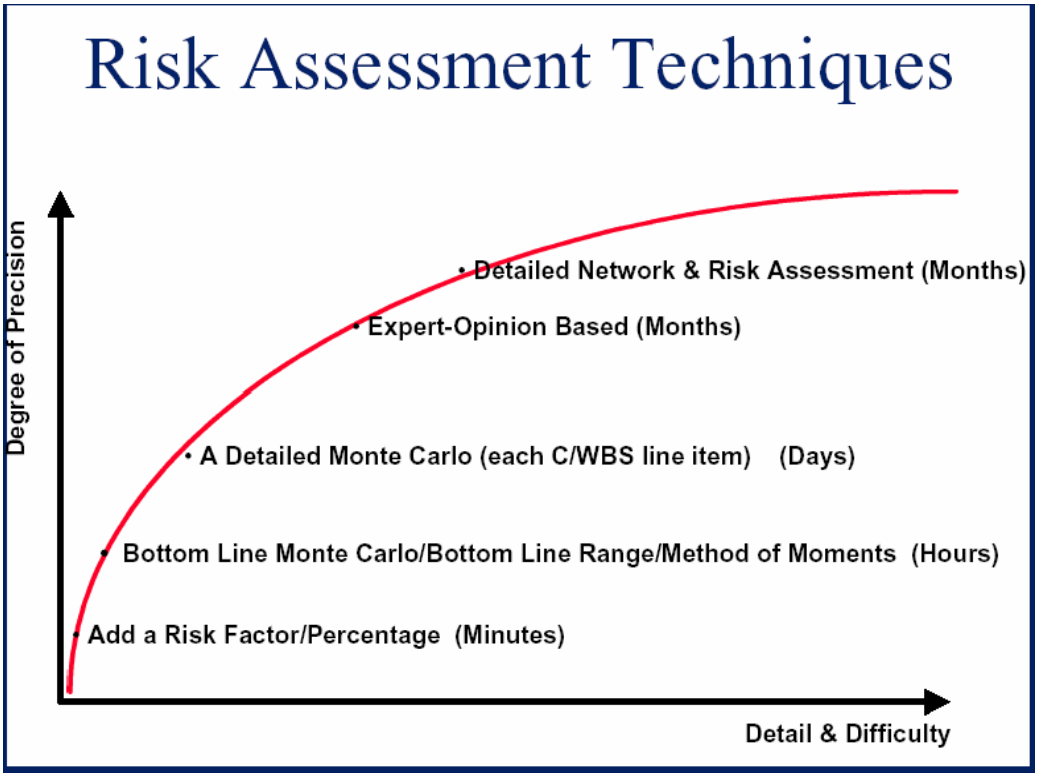


Figure 1 - Risk Assessment Techniques (Coleman, 2000:4-9)

The Detailed Network and Risk Assessment technique provides the most accurate estimates but also proves to be the most difficult, however, usually taking months to complete. This technique utilizes the Monte Carlo Simulation method and assigns various distributions to the cost or duration of each Work Breakdown Structure (WBS) item. It then utilizes a stochastic model to estimate the risk associated with each item. Estimators then accumulate this information to calculate the overall cost. While it provides the most accurate estimate, this technique also requires the most effort and takes the most time to complete (Coleman, 2000: 9).

The Expert-Opinion Based method requires conducting surveys or interviews of technical experts in the field to be estimated. Analysts use the results of these surveys

and interviews to determine the most likely, as well as the lowest and highest probable costs for each WBS item. Estimators then convert these costs to triangular distributions and conduct a Monte Carlo simulation employing these distributions. The reliability of this technique depends on the ability of those being interviewed to be subjective and accurate with their cost guesstimates (Coleman, 2000: 12).

One of the more commonly used methods, the Detailed Monte Carlo method also relies on the Monte Carlo Simulation, although it differs in that it uses historical data to build the probability distributions for each item. This method proves to be faster since the historical data can readily be found in the SARs and can be compiled quickly and used in the simulation process. The applicability and currency of the database being used present potential drawbacks associated with this method (Coleman, 2000: 16).

At the next level, three techniques are grouped: Bottom Line Monte Carlo, Bottom Line Range, and Method of Moments. Estimators use these techniques when an estimate needs to be produced in a matter of hours. Generally, the estimator uses the best alternative available to include Monte Carlo using higher level WBS distributions, a limited database, or expert opinion on a higher WBS level. Lastly, the Risk Factor method relies exclusively on available technical expertise to produce a risk factor to be used for the estimate. This method can take as little as a few minutes and analysts generally use this technique to produce a ballpark estimate (Coleman, 2000: 4).

Past Research in Cost Growth

Sipple Study (2002)

Researchers have conducted numerous studies over the years in attempts to gain a better understanding of the root causes and predictors of cost growth. Sipple (2002) provides us with a comprehensive and perceptive analysis of such studies conducted over the last twenty years. While these studies have many similarities and contain valuable insight into understanding cost growth within the DoD, none of these preceding studies combine the scope and methodology set forth by Sipple. For a thorough and in-depth look at the studies listed in Table 1, refer to Sipple (2002).

Table 1 - Previous Research (Sipple, 2002)

Cost Growth Studies
Woodward (1983)
Obringer (1988)
Singleton (1991)
Wilson (1992)
RAND – Drezner (1993)
Terry/Vanderburgh (1993)
Institute for Defense Analysis (1994)
BMDO (2000)
Christensen/Templin (2000)
Eskew (2000)
NAVAIR (2001)
RAND – Birkler (2001)

Like many of the previous studies, Sipple uses the SAR database to obtain the data necessary to conduct his study. Sipple collects data on 115 DoD programs reported between the years of 1990 and 2000. From this information, Sipple extracts 78 potential predictor variables which may be helpful in predicting future cost growth.

Sipple uses logistic and multiple regression to construct two predictive models designed to predict cost growth in the EMD phase directly attributed to engineering changes. Using a two-step approach, Sipple first utilizes logistic regression to determine whether or not a program will incur cost growth. Given that a program will experience cost growth, Sipple uses multiple regression to determine the amount of cost growth that will be realized.

Sipple's results are promising. Using logistic regression, Sipple constructs a seven-variable model to predict whether or not a program will experience cost growth. Validating this model with randomly selected data points, Sipple's model accurately predicts cost growth 69% of the time. Sipple then uses multiple regression to create a three-variable model to predict the amount of cost growth a program will experience. Again using randomly selected data points for validation, Sipple's multiple regression model accurately predicts the amount of cost growth 69% of the time with a 90% confidence bound. Sipple considers the two-step approach successful and opens the door for others to follow.

Bielecki Study (2003)

Building on the Sipple study, Bielecki (2003) incorporates the two-step approach set forth by Sipple and conducts further research using logistic and multiple regression. Bielecki also conducts his research to "focus on cost growth in the Research and

Development, Test and Evaluation (RDT&E) accounts during the Engineering and Manufacturing Development (EMD) phase of acquisition" (Bielecki, 2003: 4).

As previously mentioned, the SARs separate cost variance into seven categories: Economic, Quantity, Estimating, Engineering, Schedule, Support, and Other (Calcutt, 1993: 4). Bielecki excludes Economic and Quantity cost variances since these variances "are usually beyond the control of the cost estimator" (Bielecki, 2003: 4). Bielecki further excludes cost variances due to engineering changes since these variances have previously been examined in the Sipple study. Therefore, Bielecki seeks to build predictive regression models based on the four remaining categories: Estimating, Schedule, Support and Other.

Bielecki also finds the results to be promising. Using the same validation methodology employed by Sipple, Bielecki validates two logistic regression models for the Schedule and Estimating categories and the models validate at 85% and 78% respectively. Bielecki notes that the Support and Other categories are too sparsely populated in the database to accomplish any significant statistical regression and therefore narrows the scope of the study to Schedule and Estimating. Bielecki then validates two multiple regression models to predict the amount of cost growth due to Schedule and Estimating changes. For the Schedule category, Bielecki produces a four-variable model that validates at an 80% success rate with a 90% confidence bound. For the Estimating category, Bielecki validates his five-variable model and accurately predicts all 13 of the usable data points for a 100% accuracy rate. Again, this research is promising and warrants further exploration.

During this same time frame, similar research using the Sipple methodology was conducted by Moore (2003). Moore also used the two-step approach to predict cost growth during the EMD phase of acquisitions. Moore's research differs from Sipple and Bielecki in that it focuses on Procurement dollars in the EMD phase and not on RDT&E dollars. Because of this difference, our research focuses its attention on the research efforts of Sipple and Bielecki.

Chapter Summary

This chapter discusses the differences between cost risk and cost growth as well as the many different methods for assessing and accounting for cost risk in a program. It also references the Sipple study for its comprehensive review of cost growth studies completed over the last twenty years. Lastly, it specifically discusses the previous studies conducted by Sipple (2002) and Bielecki (2003) as we take a similar approach to conduct our research. We proceed to the next chapter which highlights our methodology that builds upon the technique set forth by Sipple (2002).

III. Methodology

Chapter Overview

This chapter describes the procedures we use to conduct this study. We first illustrate our use of the SAR database and present the methodology we utilize to obtain the data. Next, we discuss the research of Sipple (2002) as it provides the foundation for our research. We follow with the predictor variables that we extract from the SAR database to be used for our model development. Lastly, we describe our use of logistic and multiple regression to predict cost growth using our two-step approach.

SAR Database

To obtain data for our study, we use the Selected Acquisition Report database. The SARs provide the most complete, detailed source of information relating to cost variances and other pertinent program information. Each program reported in the SAR database includes explicit information required by the Office of the Under Secretary of Defense for Acquisition and Technology. These requirements allow us to populate the database with predictor variables to produce statistically sound regression models.

While the SARs provide cost variances in base-year as well as then-year dollars, we use base-year dollar figures for our research in order to exclude the effects of estimated inflation in our results. We then convert these base-year figures into a common base-year to facilitate drawing comparisons between programs. We choose to convert program figures into base-year 2002 dollars so that our results can be evaluated in terms of today's dollars. The SAR records cost variances in seven distinct categories as well as providing the total cost variance, or the sum of the following cost variance categories:

- Economic: changes in price levels due to the state of the national economy
- Quantity: changes in the number of units procured
- Estimating: changes due to refinement of estimates
- Engineering: changes due to physical alteration
- Schedule: changes due to program slip/acceleration
- Support: changes associated with support equipment
- Other: changes due to unforeseen events (Drezner, 1993:7)

As mentioned previously, since this research intends to combine the efforts of Sipple and Bielecki, it focuses only on cost growth associated with estimating, engineering, schedule, support and other. We exclude cost variances due to economic and quantity changes since cost estimators cannot reasonably account for changes due to these categories. However, because economic and quantity changes can have a significant impact on cost growth, cost analysts must normalize their estimates accordingly once these changes have taken place. Since these changes cannot be predicted and are out of the control of the cost estimator, we exclude them from this research.

The SARs present three different baseline estimates from which to calculate cost growth for this research; the planning estimate (PE), the development estimate (DE), and the production estimate (PdE). Naturally, cost growth figures vary depending on which estimate the researcher utilizes since estimates tend to be more accurate as a program matures. As programs mature, estimators update previously unknown cost figures with actual incurred costs. This increases the accuracy of the estimate and therefore reduces

overall cost growth. Since our research intends to predict cost growth in the RDT&E budget during the EMD phase of acquisition, we populate our database with only those programs using a DE baseline estimate.

The SAR database provides information related to cost, schedule and performance for all ACAT 1C and D programs from each of the military services (Knoche, 2001:1). As a result, the SAR provides a detailed, complete database consisting of those programs with a high level of government oversight. We do, however, exclude from our database any information that contains a security classification for security reasons. Despite these exclusions, our database provides a representative collection of cost, schedule and performance data for most high level programs within DoD.

Previous researchers have recognized the use of SAR data as the leading choice when compiling information on cost growth. RAND, a prominent leader in the research and development field, conducted a study on cost growth in 1993 and compiled a database using the SAR's. The RAND database, however, does not divide the cost growth into the seven categories previously mentioned. For this reason, we use the RAND database only as a guide in building our database.

Using the RAND study as a guideline, Sipple (2002) also utilizes the SARs to compile a spreadsheet database containing information related to cost, schedule and performance for each of the programs reported in the SARs. Sipple's database compiles information on programs reported in the SARs from 1990-2000 and breaks cost growth down into the seven cost growth categories. Specifically, Sipple uses logistic and multiple regression to predict cost growth caused by engineering changes during the RDT&E phase of EMD.

In 2003, Bielecki conducts a follow-on study to Sipple using a similar database. Bielecki updates Sipple's database to include the most recent current estimates and also to include any programs added to the SAR during 2001. Bielecki adopts Sipple's methodology using logistic and multiple regression to predict cost growth caused by Estimating, Scheduling, Support, and Other changes during the RDT&E phase of EMD. Bielecki also excludes economic and quantity changes for the reasons that have been discussed and he excludes engineering changes since these have been addressed by Sipple.

For this research, we continue the methodology set forth by Sipple and Bielecki and combine their efforts to predict cost growth caused by estimating, engineering, scheduling, support, and other changes during the RDT&E phase of EMD. Again, we do not consider economic and quantity cost variances since these categories, by convention, usually extend beyond the control of the cost estimator (Bielecki, 2003; 4). Further, for our research we update the database to include the most recent current estimates and include any programs added to the SAR during 2002.

SAR Database Limitations

Although the SAR database has been established as a reliable and comprehensive source of information relating to cost growth, analysts must also understand the limitations that come from using the SAR database. While none of these limitations preclude us from using the database, researchers must at least be cognizant of the limitations and their potential implications. A thorough understanding of the limitations associated with using the SAR data is important in correctly interpreting the results

attained through use of the SAR database (Drezner, 1993; 9-10). Among some of the well-known limitations of the SAR database are:

1. High level of aggregation
2. Changing baseline estimates and program restructuring
3. Changing preparation guidelines and thresholds
4. Inconsistent allocation of cost variances
5. Emphasis on effects, not causes
6. Incomplete coverage of program costs
7. Unknown and varied budget levels for program risk

A related study conducted by P. G. Hough in 1992 fully describes these and other more subtle problems associated with the SAR database.

Data Collection

Since our research intends to further explore the research conducted by Sipple (2002) and Bielecki (2003), our starting point for data collection begins where they left off. The database, created by Sipple and updated by Bielecki, includes program information on all programs required to report SARs from 1990-2001. The exceptions being those programs that contain classified information or use an estimate other than the development estimate as the baseline (Sipple, 2002:57).

Sipple initially created the database by extracting program information from the latest SAR reported for each program starting with the December 2000 SARs and working backward in time. By using only the latest SAR for each program, Sipple ensured that each data point remained independent (Sipple, 2002:57). By the time he

reached 1990, Sipple had populated the database with program information on over 110 different programs, more than enough data points to perform statistically sound regression analysis.

In 2003, Bielecki updated the database to reflect current information as of December 2001. To accomplish this, Bielecki begins by updating program information based on the latest, December 2001 SARs. Bielecki also adds seven new data points to the database. These new data points consist of programs that meet the SAR reporting criteria and have also matured at least three years into the EMD phase of acquisition.

To modify the database for our research, we update the current program information to reflect that contained in the SAR reports dated December 2002. In addition, we review the SAR database to include any new programs that have been added to the database in the previous twelve months. In doing so, we ensure that our database includes every program that currently meets our selection criteria and we guarantee that our database consists of the most recent program information available. Further, the program information that we extract from the SARs emulates that previously set forth by Sipple and consequently imitated by Bielecki. This research dictates that the methodology mirrors that of Sipple and Bielecki given that it intends to combine the efforts of these two studies and predict cost variances due to Engineering, Estimating, Schedule, Support and Other cost growth categories.

Identifying Predictor Variables

From the Sipple study, we inherit 78 potential predictor variables with which we try to predict cost growth. In addition to these variables, we seek to incorporate any

additional variables that may contain a predictive link to our response variable. When considering additional predictor variables, we must keep a few thoughts in mind:

First, the variable must be accessible to the cost estimator prior to the requirement for the development estimate. This seems logical since the estimator needs to obtain that information in order to utilize our regression model should that particular variable make it through to the final model. In the event the estimator cannot arrive at a value for that variable, he or she cannot use our model to produce an estimate and our model is therefore rendered useless (Sipple 2002:47).

Next, any variable under consideration must possess a reasonable and logical relationship with the response variable. Failure to meet this criterion could result in potential trouble. First, the estimator must understand the relationship between the variables or risk losing faith in the model and its results. Second, the relationship must be evident in the event that the results fall under executive scrutiny for similar reasons (Sipple, 2002:48).

Preliminary Data Analysis

As a rule, multiple regression requires the response variable to be from a continuous distribution. However, like Sipple and Bielecki, we find that our cost growth data comes from a mixed distribution. While roughly half of the distribution appears continuous, the other half is centered on zero, or no cost growth. This mixed distribution dictates that we split the data into two separate sets, a discrete set and a continuous set.

To convert the data set into a discrete distribution, we transform all negative cost growth to a zero and all positive cost growth to a one. For the resulting distribution, we

use logistic regression to develop a model to predict whether or not a program will experience cost growth. We then produce a continuous distribution using only those programs that incurred positive cost growth. From this distribution, we use multiple regression to develop a model to predict the amount of cost growth that the estimator should expect.

For sensitivity analysis and validation purposes, we remove 20 percent of our data before beginning our model building process. In order to eliminate any potential biases, we use the random number generator in JMP® 5.0 (SAS Institute, 2003) to assign a random value to each variable. We sort the data according to these random values and remove the bottom 27 data points, or bottom 20 percent. We use the remaining 80 percent of the database for our model building and retain the 20 percent for model validation.

Response Variables

As a result of splitting the data into two individual distributions, we identify two separate response variables. The first variable represents whether or not cost growth will occur and if so, the second variable conveys the extent of cost growth that will occur. We designate the first variable as a binary variable where a '1' indicates that a program will experience cost growth and a '0' indicates that a program will not experience cost growth (Sipple, 2002:60). We name this variable *R&D (Total) Cost Growth* to reflect the fact that this research combines cost growth attributed to the Estimating, Engineering, Schedule, Support, and Other cost categories.

For the second variable, we measure cost growth as a percentage of the development estimate and call this variable *RDT&E %*. Converting this variable to a percentage facilitates comparisons between programs and allows the variable to make intuitive sense while removing the effect of program size. In doing so, the estimator provides decision-makers with a number that is easy to comprehend and appreciate.

Predictor Variables

This research makes use of the predictor variables accumulated by Sipple (2002) and subsequently used by Bielecki (2003) as well. Sipple exhausted the program information contained in the SARs to produce a wide array of predictor variables that help predict cost growth. As this research aims to create a tool for cost estimators to produce more accurate estimates, one key component of the predictor variables is that they are readily available to the estimator at the time of the estimate.

In order to better organize the predictor variables, Sipple arranges the variables into five general categories: program size, physical type of program, management characteristics, schedule characteristics, and other characteristics. Listed below are the original predictor variables as defined and categorized by Sipple:

Program Size

- *Total Cost CY $M 2003*– continuous variable which indicates the total cost of the program in CY $M 2003
- *Total Quantity* – continuous variable which indicates the total quantity of the program at the time of the SAR date; if no quantity is specified, we assume a quantity of one (or another appropriate number) unless the program was terminated
- *Prog Acq Unit Cost* – continuous variable that equals the quotient of the total cost and total quantity variables above

- *Qty during PE* – continuous variable that indicates the quantity that was estimated in the planning estimate
- *Qty planned for R&D$* – continuous variable which indicates the quantity in the baseline estimate

Physical Type of Program

- Domain of Operation Variables
  - *Air* – binary variable: 1 for yes and 0 for no; includes programs that primarily operate in the air; includes air-launched tactical missiles and strategic ground-launched or ship-launched missiles
  - *Land* – binary variable: 1 for yes and 0 for no; includes tactical ground-launched missiles; does not include strategic ground-launched missiles
  - *Space* – binary variable: 1 for yes and 0 for no; includes satellite programs and launch vehicle programs
  - *Sea* – binary variable: 1 for yes and 0 for no; includes ships and ship-borne systems other than aircraft and strategic missiles
- Function Variables
  - *Electronic* – binary variable: 1 for yes and 0 for no; includes all computer programs, communication programs, electronic warfare programs that do not fit into the other categories
  - *Helo* – binary variable: 1 for yes and 0 for no; helicopters; includes V-22 Osprey
  - *Missile* – binary variable: 1 for yes and 0 for no; includes all missiles
  - *Aircraft* – binary variable: 1 for yes and 0 for no; does not include helicopters
  - *Munitions* – binary variable: 1 for yes and 0 for no
  - *Land Vehicle* – binary variable: 1 for yes and 0 for no
  - *Ship* – binary variable: 1 for yes and 0 for no; includes all watercraft
  - *Other* – binary variable: 1 for yes and 0 for no; any program that does not fit into one of the other function variables

Management Characteristics

- Military Service Management
  - *Svs > 1* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Svs > 2* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Svs > 3* – binary variable: 1 for yes and 0 for no; number of services involved at the date of the SAR
  - *Service = Navy Only* – binary variable: 1 for yes and 0 for no
  - *Service = Joint* – binary variable: 1 for yes and 0 for no

- *Service = Army Only* – binary variable: 1 for yes and 0 for no
- *Service = AF Only* – binary variable: 1 for yes and 0 for no
- *Lead Svc = Army* – binary variable: 1 for yes and 0 for no
- *Lead Svc = Navy* – binary variable: 1 for yes and 0 for no
- *Lead Svc = DoD* – binary variable: 1 for yes and 0 for no
- *Lead Svc = AF* – binary variable: 1 for yes and 0 for no
- *AF Involvement* – binary variable: 1 for yes and 0 for no
- *N Involvement* – binary variable: 1 for yes and 0 for no
- *MC Involvement* – binary variable: 1 for yes and 0 for no
- *AR Involvement* – binary variable: 1 for yes and 0 for no

- Contractor Characteristics
  - *Lockheed-Martin* – binary variable: 1 for yes and 0 for no
  - *Northrop Grumman* – binary variable: 1 for yes and 0 for no
  - *Boeing* – binary variable: 1 for yes and 0 for no
  - *Raytheon* – binary variable: 1 for yes and 0 for no
  - *Litton* – binary variable: 1 for yes and 0 for no
  - *General Dynamics* – binary variable: 1 for yes and 0 for no
  - *No Major Defense KTR* – binary variable: 1 for yes and 0 for no; a program that does not use one of the contractors mentioned immediately above = 1
  - *More than 1 Major Defense KTR* – binary variable: 1 for yes and 0 for no; a program that includes more than one of the contractors listed above = 1
  - *Fixed-Price EMD Contract* – binary variable: 1 for yes and 0 for no

Schedule Characteristics

- RDT&E and Procurement Maturity Measures
  - *Maturity (Funding Yrs complete)* – continuous variable which indicates the total number of years completed for which the program had RDT&E or procurement funding budgeted
  - *Funding YR Total Program Length* – continuous variable which indicates the total number of years for which the program has either RDT&E funding or procurement funding budgeted
  - *Funding Yrs of R&D Completed* – continuous variable which indicates the number of years completed for which the program had RDT&E funding budgeted
  - *Funding Yrs of Prod Completed* – continuous variable which indicates the number of years completed for which the program had procurement funding budgeted
  - *Length of Prod in Funding Yrs* – continuous variable which indicates the number of years for which the program has procurement funding budgeted
  - *Length of R&D in Funding Yrs* – continuous variable which indicates the number of years for which the program has RDT&E funding budgeted

- *R&D Funding Yr Maturity %* – continuous variable which equals *Funding Yrs of R&D Completed* divided by *Length of R&D in Funding Yrs*
- *Proc Funding Yr Maturity %* – continuous variable which equals *Funding Yrs of R&D Completed* divided by *Length of Prod in Funding Yrs*
- *Total Funding Yr Maturity %* – continuous variable which equals *Maturity (Funding Yrs complete)* divided by *Funding YR Total Program Length*

- EMD Maturity Measures
  - *Maturity from MS II in mos* – continuous variable calculated by subtracting the earliest MS II date indicated from the date of the SAR
  - *Actual Length of EMD (MS III-MS II in mos)* – continuous variable calculated by subtracting the earliest MS II date from the latest MS III date indicated
  - *MS III-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD (MS III-MS II in mos)*
  - *Actual Length of EMD using IOC-MS II in mos* – continuous variable calculated by subtracting the earliest MS II date from the IOC date
  - *IOC-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD using IOC-MS II in mos*
  - *Actual Length of EMD using FUE-MS II in mos* – continuous variable calculated by subtracting the earliest MS II date from the FUE date
  - *FUE-based Maturity of EMD %* – continuous variable calculated by dividing *Maturity from MS II in mos* by *Actual Length of EMD using FUE-MS II in mos*
- Concurrency Indicators
  - *MS III Complete* – binary variable: 1 for yes and 0 for no
  - *Proc Started based on Funding Yrs* – binary variable: 1 for yes and 0 for no; if procurement funding is budgeted in the year of the SAR or before, then = 1
  - *Proc Funding before MS III* – binary variable: 1 for yes and 0 for no
  - *Concurrency Measure Interval* – continuous variable which measures the amount of testing still occurring during the production phase in months; actual IOT&E completion minus MS IIIA (Jarvaise, 1996:26)
  - *New Concurrency Measure %* – continuous variable which measures the percent of testing still occurring during the production phase; (MS IIIA minus actual IOT&E completion in moths) divided by (actual minus planned IOT&E dates) (Jarvaise, 1996:26)

Other Characteristics

- *# Product Variants in this SAR* – continuous variable which indicates the number of versions included in the EMD effort that the current SAR addresses
- *Class – S* – binary variable: 1 for yes and 0 for no; security classification Secret

- *Class – C* – binary variable: 1 for yes and 0 for no; security classification Confidential
- *Class – U* – binary variable: 1 for yes and 0 for no; security classification Unclassified
- *Class at Least S* – binary variable: 1 for yes and 0 for no; security classification is Secret or higher
- *Risk Mitigation* – binary variable: 1 for yes and 0 for no; indicates whether there was a version previous to SAR or significant pre-EMD activities
- *Versions Previous to SAR* – binary variable: 1 for yes and 0 for no; indicates whether there was a significant, relevant effort prior to the DE; a pre-EMD prototype or a previous version of the system would apply
- *Modification* – binary variable: 1 for yes and 0 for no; indicates whether the program is a modification of a previous program
- *Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the program had a prototyping effort
- *Dem/Val Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the prototyping effort occurred in the PDRR phase
- *EMD Prototype* – binary variable: 1 for yes and 0 for no; indicates whether the prototyping effort occurred in the EMD phase
- *Did it have a PE* – binary variable: 1 for yes and 0 for no; indicates whether the program had a planning estimate
- *Significant pre-EMD activity immediately prior to current version* – binary variable: 1 for yes and 0 for no; indicates whether the program had activities in the schedule at least six months prior to MSII decision
- *Did it have a MS I* – binary variable: 1 for yes and 0 for no
- *Terminated* – binary variable: 1 for yes and 0 for no; indicates if the program was terminated

After reviewing the predictor variables and updating the database to reflect current dollar figures, we decide to make some changes to these predictor variables. The changes are made for clarity purposes and to facilitate producing a statistically sound model. The following list documents these changes:

- Delete *Domain of Operation* – Air/Sea/Land/Space variables make this redundant
- Delete *Proc Cost Growth* - includes all seven categories of cost growth; only five are needed
- Delete *Class S-R* – all of our SARs are classified secret or lower, this variable duplicates *Class S*
- Delete *Is MSIII Complete?* – always zero since MSIII cannot be complete for our programs

- Delete *RAND Concurrency Measurement Interval and RAND Concurrency Measurement Interval %* - does not apply to programs in the EMD phase
- Delete *Terminated?* – this variable cannot be used if the program still operates and therefore provides no value added to our model
- Delete the following variables for lack of data points (less than 30 would remain after we remove the 20 percent validation set):
    - *FOT&E End Planned*
    - *FOT&E End current estimate*
    - *MSIIIa planned & current estimate*
    - *MSIIIb planned & current estimate*
    - *FUE planned*
    - *FUE current estimate*
    - *Maturity from MSII (current calculation in months)*
    - *Qty in PE*
- Add *LRIP Planned?* – binary with 1 for yes and 0 for no to indicate whether the program has Low Rate Initial Production
- Add *Space (RAND)* – missing from the original database, but needed for full accountability of the included programs
- Change of variable name:
    - *Qty Planned for R&D$* to *Qty Planned for R&D*
    - *Earliest Actual MSII Date* to *Current Actual MSII Date*
    - *Earliest Actual MSIII Date* to *Current Actual MSIII Date*
    - *Actual Length of EMD using (E-B)* to *Time from MSII to IOC in months*
    - *Program Acquisition Unit Cost* to *Unit Cost*
    - *Maturity of EMD using IOC* to *Maturity of EMD at IOC* (also corrected the formula so that if IOC occurs after MSIII, the percentage cannot exceed 100%)

Upon initial investigation of the contractor variables, Sipple finds that the SARs list 45 different defense contractors for our programs. In order to produce statistically significant variables, Sipple consolidates these contractor variables to reflect real-world corporate mergers within the defense industry during the 1990s (Sipple, 2002:65). The results of this consolidation are evidenced by the six contractor variables listed under the predictor variables.

Logistic Regression

As previously mentioned, our research intends to build two predictive models, the first to predict whether or not a program will experience cost growth and the second to predict the amount of cost growth that will occur. To build the first model, we utilize logistic regression. Logistic regression is used when the response variable is binary, usually either a '1' or a '0'. For our database, we assign a '1' to each program that incurs cost growth and a '0' to each program that does not incur cost growth. For the purposes of our research, negative cost growth is not considered and programs that incur negative cost growth are assigned a '0'.

We use JMP® 5.0 (SAS Institute, 2003) software to perform this logistic regression and build our model. However, unlike the step-wise regression tool available for multiple regression, logistic regression offers no automated method of running the regressions. As a result, we systematically compute thousands of regressions in order to obtain the best model possible.

We start by running each predictor variable individually and recording our results. We then rank the variables according to the resulting *p*-values and carry forward those one-variable models with a *p*-value of less than 0.05. We then combine those one-variable models with each of the other predictor variables to produce every combination of two-variable models. At this point, we decide to carry forward our top ten resulting two-variable models, again based on cumulative *p*-value, since the number of models with a cumulative *p*-value of less than 0.05 increases dramatically. We then carry those ten two-variable models forward and combine them with the other predictor variables to produce every combination of three-variable models. We repeat this process until the

result of adding an additional variable to our top model causes our cumulative *p*-value to exceed 0.10, any individual *p*-value to exceed 0.05, or causes our data point-to-variable ratio to fall below 10:1.

In order to expedite the regression process, after each round of regressions, we remove those predictor variables that fail to produce any models that meet the selection criteria mentioned above. Once we produce our final model, we combine that model with each of the predictor variables that have been removed to ensure that none of these variables could improve our model. We intend to find the best model possible that meets the above criteria and then validate our model using the 20 percent of data points that have been set aside for validation.

Multiple Regression

The next model we build aims to predict the amount of cost growth that will occur. We build this model using multiple regression. Again, we utilize JMP® 5.0 software to perform this multiple regression and to build our model. For this portion of our model building, we exclude from our database those programs that did not incur cost growth or incurred negative cost growth. We then use the remaining data points, those that do incur cost growth, to build our multiple regression models. To maintain consistency and ensure thoroughness, we employ the same methodology demonstrated during the logistic regression portion of our research.

As we did for logistic regression, we start by running each predictor variable individually and recording the resulting *p*-values. With multiple regression, however, we carry forward our top ten one-variable models instead of using a 0.05 *p*-value cutoff.

This change ensures that a significant number of one-variable models get carried forward to the next stage of model building. As with the logistic portion of our research, we continue to run each of our top ten models with the remaining predictor variables to produce every possible combination. We continue this process until the result of adding an additional variable causes our cumulative *p*-value to exceed 0.10, any individual *p*-value to exceed 0.05, or causes our data point-to-variable ratio to fall below 10:1. We again intend to discover the best model possible and then validate our model using the 20 percent of data points previously set aside for validation.

Review of Methodology

This chapter discusses the research methodology set forth in this study. We analyze the SARs as our primary source of data, as well as some of the well-known limitations in using the SAR database. We describe our data collection process and examine the predictor variables that we utilize in our model building. Finally, we discuss the need to combine both logistic and multiple regression techniques and the manner in which we do so to complete our research. Next, we introduce and analyze the results achieved via this methodology.

IV. Results

Chapter Overview

This chapter details the results of our logistic and multiple regression analysis. First, we expound on the methodology and criteria we utilize to arrive at our final models for each of the two regressions. We then test each of our final models to ensure that each is statistically valid. Next, we discuss the applicability and robustness of our models for the cost estimators in the field. Finally, we validate each regression model using the 27 data points previously set aside for model validation.

Multiple Regression

As previously discussed, multiple regression generally requires that the response variable comes from a continuous distribution. A preliminary review of our cost growth data reveals that the response variable *RDT&E %* comes from a mixed distribution, as seen in the stem and leaf plot shown in Figure 2. The plot shows a discrete mass at zero cost growth and a reasonably continuous distribution for the rest of the data. As a result of this mixed distribution, we transform the data set into two separate sets, a discrete set and a continuous set. Next, we follow the two-step methodology established by Sipple (2002). First, we use logistic regression to develop a model to predict whether or not a program will incur cost growth. We then use multiple regression to develop a model to predict how much cost growth a program will incur given that the program will incur some positive cost growth.

Stem	Leaf	Count
4	7	1
4		
3		
3		
2	77	2
2	11	2
1	56699	5
1	0011223	7
0	55555666777888899	17
0	11111111111112222222233333334444444	35
-0	44332222222211110000000000000000	32
-0	888875	6
-1		

Figure 2 - Stem and Leaf plot of *RDT&E %* (Stem = 10%, Leaf = 1%)

Logistic Regression Results

As discussed in Chapter 3, we systematically compute thousands of regressions using our transformed database to find the most predictive logistic model. As logistic regression requires the response variable to be binary, we assign a '1' to each program that incurs cost growth and a '0' to each program that does not incur cost growth. Using this new database, we build our logistic model.

We regress each predictor variable independently and document our results, giving us each possible one-variable model. Using model *p*-values as our measuring stick, we rank-order these results and carry forward our best one-variable models, those models that have a *p*-value of less than 0.05. We then combine each of those variables with the remaining 76 variables to produce every possible two-variable model combination. Again, we document our results and rank-order those results based on cumulative *p*-value. Since the number of models whose cumulative *p*-value is less than

0.05 increases dramatically, we decide to carry forward our ten best two-variable models from which we produce our three-variable models. Again, we combine our top two-variable models with each of the remaining predictor variables to produce every possible combination of three-variable models, which we then document and rank-order. We continue this process, building our models one variable at a time, until the result of adding any additional variables to our top model causes our model to have a cumulative *p*-value of greater than 0.10, any individual *p*-value of greater than 0.05, or a data point-to-variable ratio of less than 10:1.

As we progress through our model building process, we must eventually compare each of our top candidate models and identify which model ultimately represents the best model given our database. The measures that we utilize to draw comparisons between our top models are the R^2 (U), the data point-to-variable ratio, and the area under the ROC. Table 2 summarizes each of these measures results for the best model produced for each generation of models. Following is a brief description of each of these statistical measures.

Table 2 - Logistic Regression Evaluation Measures

Number of Predictors	1	2	3	4	5	6
R^2 (U)	0.1713	0.2006	0.2499	0.2912	0.3565	0.4359
Data points	108	108	108	104	97	96
Data point-to-var. ratio	108	54	36	26	19.4	16
Area under ROC	0.7861	0.8029	0.8215	0.8448	0.8820	0.9115

The first measure of comparison for our models is the R^2 (U), also known as the uncertainty coefficient. According to the JMP® help menu, the R^2 (U) is the ratio of the

difference to the reduced negative log likelihood values. More specifically, the R^2 (U) represents the difference of the negative log likelihood of the fitted model minus the negative log likelihood of the reduced model divided by the negative log likelihood of the reduced model. As with the Adjusted R^2 associated with ordinary least squares regression, the R^2 (U) ranges from 0, which indicates no predictive capability, to a 1, which indicates a perfect model fit. However, high R^2 (U) values are rare when using a nominal model (JMP® 5.0, 2002: Help). With this in mind, we select models with the highest R^2 (U) values while realizing that our expectations should be reasonable.

The next measure of comparison for our models is the data point-to-variable ratio. The data point-to-variable ratio underlies the importance of obtaining a large sample size of data. A large sample size represents a larger portion of the total population and therefore helps to avoid over fitting the model. According to Neter et al, a model should strive to contain at least ten data points for each predictor variable to avoid over fitting the model to the database. Models with a data point-to-variable ratio between 6:1 and 10:1 may be considered if the benefits of adding the additional variables can be justified but any model that falls below a 6:1 ratio should not be considered (Neter, 1996:437). For our research, we strive for and are successful at maintaining a minimum ratio of at least 10:1.

Next we take into account the area under the Receiver Operating Characteristic (ROC) curve for comparison between models. The JMP® help menu defines the ROC curve as a graphical representation of the relationship between false-positive and true positive rates. The area under the curve is a common index used to summarize the information contained in the curve. The area under the curve represents the probability

that the model will accurately predict whether or not a program will incur cost growth (JMP® 5.0, 2002: Help). As such, we choose those models that maximize the area under the ROC curve.

Lastly, we consider the *p*-values for each of the model's parameters. Each parameters *p*-value represents the statistical significance of that parameter, the lower the *p*-value the greater the significance. While we prefer the *p*-values be as low as possible, we reject any model that contains a parameter with a *p*-value greater than 0.05 to avoid over-fitting the model to the fitted data instead of the overall population (Sipple, 2002: 79). For each of our top models listed in Table 2, we experience no parameter *p*-values greater than 0.05. However, as we continue our model building process to produce all combinations of seven-variable models, we find that we produce no seven-variable models that do not significantly surpass our *p*-value ceiling of 0.05. For this reason, we choose our top six-variable model as our best model up to this point and decide to test this model to see if it can be improved through higher-order terms or interactions.

We conduct testing for higher-order terms by taking each of the six predictor variables in our model and squaring, then cubing each variable to see if this results in a significant improvement in our performance measures. During this testing, we find that squaring the variable *R&D Funding Yr Maturity %* produces significant gains in both the R^2 (U) value and the area under the ROC curve, while also reducing the cumulative parameter *p*-values for our model. As a result, we proceed to interaction testing using this new six-variable model.

Interaction testing is accomplished by multiplying each of the six variables in our model by each of the other five remaining variables. This results in an additional variable

called an interaction variable. During interaction testing, we discover that multiplying *Maturity (Funding yrs complete)* by *EMD Prototype* also significantly increases our R^2 (U) value and the area under the ROC curve. We find that the interaction term slightly increases our cumulative parameter *p*-values but does not cause any individual *p*-value to exceed our ceiling of 0.05, nor does the addition of a seventh variable cause us to drop below our optimal data point-to-variable of 10:1. Table 3 shows the changes in parameter *p*-values as we proceed through this testing phase.

Table 3 - Changes in Parameter *P*-values

Predictor Variable	Orig. 6 variables	6 w/ higher-order	7 w/ interaction
Svs > 3	0.0119	0.0144	0.0123
Maturity (Funding Yrs)	0.0003	0.0005	0.0006
R&D Funding Yr Maturity %	0.0010		
Risk Mitigation?	0.0114	0.0081	0.0026
EMD Prototype?	0.0114	0.0103	0.0318
Program have a MS I?	0.0054	0.0045	0.0020
R&D Funding Yr Maturity % - squared		0.0004	0.0004
Maturity (Funding Yrs) * EMD Prototype?			0.0072
Cumulative *p*-value	0.0414	0.0382	0.0569

Consequently, we proceed using our new seven-variable model, to include our new interaction term. For the rest of our interaction testing, we find no further interactions that improve upon our current model. Lastly, we test the inverse, natural log and exponents for each variable currently in our model to see if we can improve our model further. We find that none of this further testing improves our seven-variable model. Table 4 shows the improvements in our evaluation measures gained by

substituting our higher-order term and including our interaction variable in our final model.

Table 4 - Improvements with Higher-order Term and Interaction

Number of Predictors	Orig. 6 variables	6 w/ higher-order	7 w/ interaction
R^2 (U)	0.4359	0.4581	0.5357
Data points	96	96	96
Data point-to-var. ratio	16	16	13.7
Area under ROC	0.9115	0.9149	0.9344

With the testing complete and based on our aforementioned model criteria, we conclude that our current seven-variable model provides the most predictive capability to determine whether or not a program will incur cost growth (Appendix A). We now proceed to the validation process.

Logistic Regression Validation

To validate our model, we use 27 data points that we select randomly from the original data set containing 135 data points. From these 27 selected data points, we find that 6 data points cannot be used for validation as these points having missing values for at least one of the variables in our final model. The remaining 21 data points represent approximately 18% of the 117 useable data points from our entire data set. While this falls slightly short of our original validation goal of 20%, we conclude that the difference is minimal and proceed with the validation.

To validate the remaining 21 data points, we save the functionally predicted values generated in JMP® for each of the data points. For each data point, JMP® assesses the probability of that program incurring cost growth. JMP® then assigns a '1'

to any point whose probability is 0.5 or greater and a '0' to all other points (Sipple, 2002: 82). We then compare these predicted outcomes with the actual outcomes for each program. We find that our model accurately predicts 15 out of the 21 data points for a success rate of 71%. We are satisfied with these results and feel that this model has good predictive capability as well as being applicable in the field.

Multiple Regression Results

We now proceed with step two of our two-step methodology, building a predictive model to forecast the amount of cost growth expected given that the program will incur cost growth. For this phase of model building we remove from our randomly selected data set those programs that experience either negative or no cost growth. We exclude these data points to improve our model's accuracy by preventing data points outside our range of interest from skewing our results (Sipple, 2002: 83). We find that this action excludes 31 of our original 108 data points, leaving us with 77 data points from which we build our predictive model. We construct this model utilizing the same 77 predictor variables as our logistic model. However, for this phase of model building we change our response variable to *RDT&E %*, which calculates the percent increase of cost growth from the DE baseline estimate.

Preliminary analysis indicates that our response variable does not approximate a normal distribution. We anticipate this scenario since earlier research conducted by Sipple (2002) and Bielecki (2003) met with similar circumstances. Like Sipple and Bielecki, we find that a natural log transformation of the response variable successfully normalizes our distribution and also corrects for non-constant variance among the

residuals that we find when the response variable is not transformed. Figure 3 shows the results of this natural log transformation using JMP®,

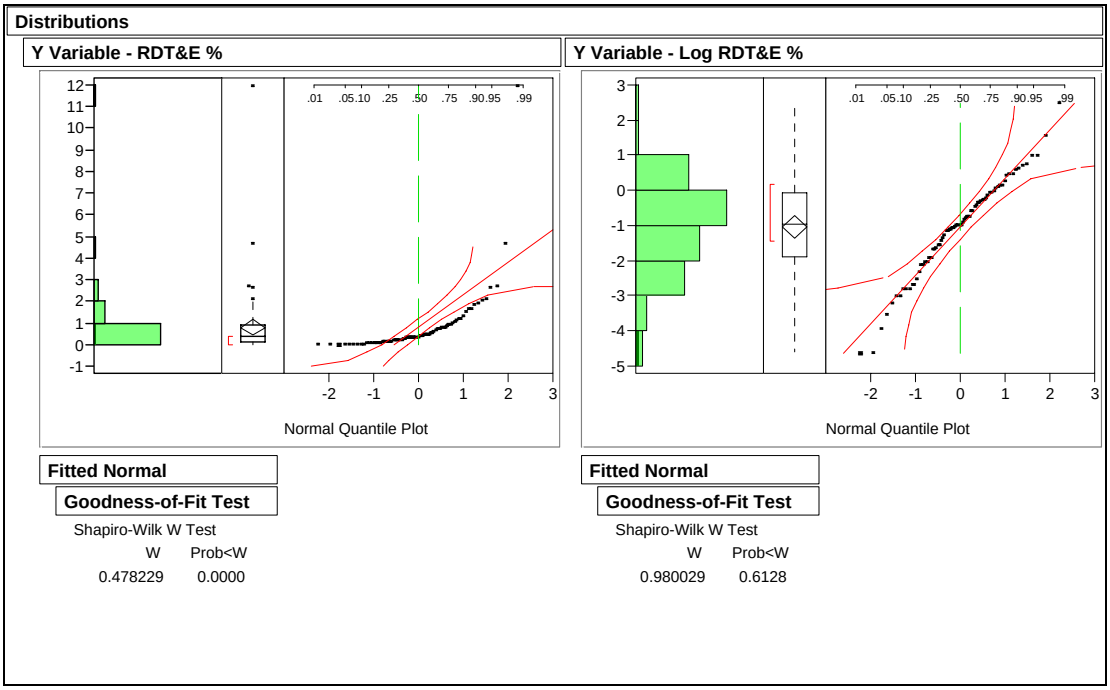


Figure 3 - Distribution of *RDT&E %* and Log *RDT&E %*

To maintain consistency and ensure thoroughness, we employ the same model building methodology as demonstrated in the logistic regression segment of our research. We run each predictor variable individually and record the resulting *p*-values. At this point, we carry forward our ten best one-variable models. We refrain from using a 0.05 *p*-value cutoff to ensure a significant number of one-variable models get carried forward to build the next generation of models. Again we continue this process and run each of our top models with each of the remaining predictor variables to produce every possible combination and then rank the results. We repeat this process until the result of adding any additional variables causes our cumulative *p*-value to exceed 0.10, any individual *p*-values to exceed 0.05, or causes our data point-to-variable ratio to fall below 10:1.

The evaluation measures for the multiple regression model building are comparable to those for logistic regression except that multiple regression focuses on the Adjusted R^2 instead of the R^2 (U). We prefer the Adjusted R^2 to the regular R^2 since the Adjusted R^2 protects against artificial inflation due to simply adding additional variables to our model. We continue to track cumulative *p*-values and data point-to-variable ratios for selection criteria of our most predictive model. Table 5 summarizes each of these measures results for the best model produced for each generation of models.

Table 5 - Multiple Regression Evaluation Measures

Number of Predictors	1	2	3	4	5	6
Adjusted R^2	0.0559	0.1342	0.2201	0.2905	0.3295	0.3620
Data points	75	72	72	72	72	60
Data point-to-var. ratio	75	36	24	18	14.4	10
Cum. *p*-value	0.0231	0.0105	0.0070	0.0095	0.0418	0.1036

From Table 5, we see that we experience a significant increase in the Adjusted R^2 for each generation of models when an additional variable is added. Further, our data point-to-variable ratio remains at or above our desired 10:1 ratio up to our six-variable model. We do find cause for concern as the cumulative *p*-value for our six-variable model exceeds 0.10. However, after examining the model we find that no individual *p*-values within the model exceed our limit of 0.05. For this reason, we decide to allow this minor deviation and accept our six-variable model as a viable model. As a result of our cumulative *p*-value exceeding 0.10 and since adding any additional variables would cause our data point-to-variable ratio to fall below 10:1, we decide to discontinue our model building at this point and determine that, based on the prescribed evaluation measures,

our current six-variable model provides the most predictive capability to determine the amount of cost growth that a program will incur. We test for and conclude that this model passes the statistical assumption tests of normality and constant variance at an alpha of 0.05. Since we exclude all dependent programs from our original data set and there is no obvious serial correlation, we assume independence of the residuals. Lastly, we test for multicollinearity by checking that all variance inflation factors are less than ten (Neter, 1996:387) and find all variance inflation factors at an acceptable level.

As with logistic regression, we now conduct higher-order term and interaction testing on our top model to test if the model can be improved upon. We find that our model is not improved by substituting higher-order terms or by adding interaction variables. Lastly, we test the inverse, natural log and exponents for each variable in our model and find again that no significant gains are realized. With the testing complete and based on our evaluation measures, we conclude that our current six-variable model remains the most predictive model to determine whether or not a program will incur cost growth (Appendix B). We now proceed to the validation process.

Multiple Regression Validation

To validate our multiple regression model, we use the same randomly selected 27 data points that were used during logistic regression validation. Upon initial review of these data points, we find that 11 data points do not incur cost growth, leaving 16 data points from which to validate our model. We use 11 of these 16 data points as 5 data points are lost due to missing values. These 11 data points represent approximately 12% of the 94 programs in our data set that incur cost growth. As with our logistic regression

validation, this falls short of our original validation goal of 20%, but we again proceed with the validation.

To validate the remaining 11 data points, we combine the validation data set with our original data set and save the predicted values for each model to be validated. We then create a 90 percent upper prediction bound and transform our response variable back to normal. We measure the success of our model by determining whether the actual percentage of cost growth incurred is captured within our 90 percent prediction bound. If the actual percentage of cost growth is less than the prediction bound, it is determined to be a successful prediction. From our validation, we find that our model accurately predicts 10 out of the 11 data points at a prediction bound of 90 percent for a success rate of 91%. We are obviously pleased with these results and feel that this model has significant predictive capability.

Chapter Summary

This chapter reviews the methodology utilized to obtain our most predictive regression models, discusses our criteria for choosing those models and analyzes the subsequent validation results. We determine that both final models perform reasonably well during the validation phase and both are fairly universal in their applicability. In the next and final chapter, we discuss our conclusions based on this research, compare these results with previous research and address the potential for real-world application.

V. Conclusions

Chapter Overview

This chapter recapitulates the concern regarding cost growth in DoD acquisitions and how that concern provides the impetus for this research. We then reconsider preceding cost growth research presented in the literature review. Next, we briefly review the methodology utilized for this research and discuss the results obtained using this methodology. Lastly, as our research is follow-on in nature, we compare our results with the results achieved from prior research as well as offer recommendations for future research.

Explanation of the Issues

Problems associated with cost growth in major weapon systems procurement have plagued the DoD for over three decades. The inability of the cost estimating community to provide timely and accurate cost estimates affects Congress' ability to draw accurate comparisons between competing weapons systems. When these cost estimates are inaccurate, they can potentially have a negative influence on congressional decisions regarding the allocation of tax money provided in good faith by the American public.

The objective of this research is to reduce cost growth by providing the cost estimating community with a predictive tool to account for program risk early in a program life-cycle and improve the accuracy of the development estimate. We accomplish this objective using a two-step methodology established by Sipple (2002) and subsequently employed by Bielecki (2003). We use logistic regression to determine

whether or not a program will experience cost growth and if so, we then use multiple regression to determine the amount of cost growth that should be expected.

Review of Literature

We conduct a thorough review of literature pertaining to weapon systems cost growth within the DoD. We find that the SAR database provides the most complete, detailed source of information relating to cost variances and other pertinent program information. We also find that although many previous cost growth studies contain similarities and share common traits with our research, only two previous studies compare with our research in both methodology and scope, Sipple (2002) and Bielecki (2003). Sipple establishes the two-step methodology to predict cost growth and focuses his research on cost growth of RDT&E dollars due to engineering changes during the EMD phase of acquisition. In addition, Sipple compiles a substantial collection of predictor variables from which to build a predictive model. As follow on, Bielecki (2003) adopts the same methodology to predict cost growth of RDT&E dollars due to schedule, estimating, support, and other changes during the EMD phase of acquisition. Our study builds upon this line of research and combines the efforts of Sipple and Bielecki, that is, we seek to predict cost growth of RDT&E dollars due to engineering, schedule, estimating, support and other changes during the EMD phase of acquisition. We exclude the two remaining cost variance categories economic and quantity as these categories are usually beyond the control of the cost estimator. As a result of combining these two studies, we conduct our research using the same predictor variables and methodology set forth by Sipple and Bielecki with only minor deviations.

Review of Methodology

For our research, we expand upon the SAR database established by Sipple (2002) and subsequently revised by Bielecki (2003). The database contains program information on all major acquisition programs, from 1990 through 2001, that use the development estimate as the baseline estimate. For our study, we update this database to reflect the most recent program information available and to include any additional programs that become eligible for our database during the 2002 calendar year. As a result, our database contains information on all major acquisitions programs, from 1990 through 2002, that use the development estimate as the baseline estimate. Our complete database now consists of 135 data points, 108 of which we use for model building and 27 randomly selected data points that we reserve for validation. We eliminate the effects of inflation by converting all dollar amounts into base year 2002 dollars.

Next, we compute the response variables for both the logistic and multiple regression portions of our study. For logistic regression, the response variable *RDT&E Cost Growth?* is binary. We assign a '1' to those programs that experience cost growth and a '0' to those programs that experience no cost growth or negative cost growth. We then divide the total cost variance by the baseline cost for each program to compute our multiple regression response variable, *RDT&E %*. We find that a natural log transformation of this response variable successfully normalizes our distribution and corrects non-constant variance among the residuals.

We use JMP® to systematically compute thousands of regressions in order to construct the most predictive models possible. We continue building our models one variable at a time until the addition of another variable results in our model exceeding our

evaluation measure criteria. We then use those evaluation measures to compare our top models and select our most predictive model for each regression technique. Finally, we substantiate our top models using those data points previously set aside for validation.

Restatement of Results

For the logistic regression portion of our research, we find that a seven-variable model produces the best results. We discover that substituting a higher-order term and the addition of an interaction term improves the predictive capability of our model. Upon validation, our model accurately predicts cost growth in 15 out of 21 programs for a success rate of 71%.

For the multiple regression segment of our study, we find a six-variable model produces the best results. We test for higher-order terms and interactions but find no significant improvements on our current model. Using a 90 percent prediction bound, our model accurately predicts the amount of cost growth in 10 out of 11 programs for a success rate of 91%. We are pleased with the results of each model and conclude that both models exhibit significant predictive capability.

Comparison with Previous Research

As we conclude our research, we compare the results of our models that predict overall cost growth of RDT&E dollars with the results of the models produced by Sipple (2002) and Bielecki (2003) that predict cost growth due to engineering, estimating, and schedule changes. We draw these comparisons to determine if the models are more predictive when we isolate cost growth by cost variance category or combine categories to represent the phase of acquisition that incurs the cost growth. We also evaluate the

predictor variables present in each of the final models as we seek to identify patterns or key predictor variables that identify cost growth. As previously mentioned, economic and quantity cost variances are not considered in these studies since these categories are usually beyond the control of the cost estimator. Furthermore, cost variances resulting from support or other changes are not researched individually due to an insufficient number of data points.

First, we compare the evaluation measures and validation results of the logistic regression models. Table 6 displays these measure statistics and validation results for each of the four logistic models. Availability percent represents the percent of total validation points available that each model is able to validate.

Table 6 – Logistic Regression Model Comparison

Cost Category	R^2 (U)	ROC	Ratio	Availability %	Validation %
Engineering - Sipple	0.6012	0.9481	8.7	52%	69%
Estimating - Bielecki	0.4184	0.8981	12.6	92%	78%
Schedule - Bielecki	0.4808	0.9200	8.8	28%	86%
RDT&E - Genest	0.5357	0.9344	13.7	78%	71%

Judging by the R^2 (U) and the ROC, it appears that the engineering model holds a slight predictive advantage over the remaining three models. However, we notice that the data point-to-variable ratio falls into the cautionary zone below 10:1 and may be the result of over-fitting the model to the data available. These concerns are further heightened by the low availability percentage and validation percentage. As a result, we hold short of declaring this the most predictive model. While the schedule model presents the highest validation percent, we are again concerned about the data point-to-

variable ratio and the extremely low availability percent. Further review of the models results in no major differentiation between the four. We conclude that for the logistic regression portion of this research, there is no significant advantage gained by either isolating each cost variance category individually or by combining these categories.

Next, we assess the evaluation measures and validation results for the multiple regression models. Table 7 shows the statistics and validation results for each of the four multiple models.

Table 7 – Multiple Regression Model Comparison

Cost Category	Adj R^2	Ratio	Availability %	Validation %
Engineering - Sipple	0.4222	14.0	93%	69%
Estimating - Bielecki	0.5225	8.8	87%	100%
Schedule - Bielecki	0.6190	9.0	91%	80%
RDT&E - Genest	0.3620	10.0	69%	91%

Similar to the logistic regression models, we find that the model with the highest Adjusted R^2, the schedule model, does not result in the highest validation percent. This may also be a result of the data point-to-variable ratio falling into the cautionary zone and the model over-fitting the data. The combined RDT&E model results in a significantly lower Adjusted R^2 than the other models but results in a surprising second-best validation percent. We again deduce that there is no significant benefit gained by isolating each cost variance category individually or by combining these categories for the multiple regression portion of our research.

We now compare predictor variables for each of the logistic regression models to ascertain any trends or key predictor variables in predicting cost growth. Table 8 lists the predictor variables found in each of the final logistic models.

Table 8 – Logistic Regression Predictor Variable Comparison

Engineering - Sipple	Estimating - Bielecki	Schedule - Bielecki	RDT&E - Genest
Actual Length of EMD	Length of R&D in Funding Yrs	Maturity (Funding Yrs complete)	Svc > 3
MSIII-based Maturity of EMD %	Versions Previous to SAR	Army Involvement	Maturity (Funding Yrs complete)
Modification	Navy Involvement	Versions Previous to SAR	R&D Funding Yr Maturity %
Length of R&D in Funding Yrs	PE	Prototype	Risk Mitigation
Length of Prod in Funding Yrs	Lead Svc = DoD	Northrop Grumman	EMD Prototype
Actual Length of EMD (IOC-MSII)	Program have a MS I		Program have a MS I
Land Vehicle	Prototype		

From this review, we discover a handful of predictor variables that appear in two of the four models. However, we do not uncover any unanimous variables or revealing trends that lead us to draw any conclusions for more predictive models in the future. We proceed to the multiple regression models and to Table 9 which identifies the predictor variables found in each of the final multiple models.

Table 9 –Multiple Regression Predictor Variable Comparison

Engineering - Sipple	Estimating - Bielecki	Schedule - Bielecki	RDT&E - Genest
Maturity from MS II	IOC-based Maturity of EMD %	Boeing	Northrop Grumman
No Major Def Contractor	Proc Funding Yr Maturity %	Land Vehicle	Funding Yrs of R&D Completed
Prog Acq Unit Cost	General Dynamics	Lead Svc = Navy	Maturity of EMD at IOC %
	Lead Svc = Navy	Program have a MS I	Prototype
	PE		Significant pre-EMD activity
			LRIP Planned

Review of the multiple regression models reveals similar results. We do not find any common variables between the four models nor do we expose any trend to shed light on future cost growth research.

Comparison of these models, predictor variables, and validation results reveals no considerable advantage realized from one model to the next. However, each model provides a statistically sound predictive model to be used to predict cost growth. We therefore encourage that each of these models be taken into consideration for use, ultimately selecting the model that best fits the needs of the cost estimator.

Recommendations

As do previous studies using this two-step approach, our research concludes that the use of logistic regression is warranted and in fact, preferred. Logistic regression allows the cost estimator to determine whether or not a program will incur cost growth, potentially saving the estimator a significant amount of time if the answer is no. If the answer is yes, this two-step method offers a more reliable depiction as to the amount of cost growth to be expected as it prevents those programs that do not incur cost growth from skewing the results. Furthermore, logistic regression allows the cost estimator the opportunity to adjust the level of certainty for the predicted outcome. For this study, we use a cut-off value of 0.50 to assign a '1' or a '0' to each data point. However, estimators may adjust this cut-off point in either direction to produce a more or less conservative outcome. This flexibility provides the estimator the capability to conduct sensitivity analysis for each result. Lastly, once positive cost growth is predicted, the multiple

regression model provides a tool that is statistically sound and allows the estimator to adjust the upper prediction bound accordingly based on mission needs.

Possible Follow-on Theses

We find that the two-step methodology presents a valuable tool providing significant predictive capability and therefore support further use of this methodology for future cost growth research. Furthermore, we encourage continued use of the extensive database produced from this line of research as we could find no other database that provided such a comprehensive overview of so many programs. Potential areas for further research include, but are not limited to:

- Isolate programs that did not have significant cost overruns and evaluate their risk estimating methodology to determine if there is a best methodology (Sipple, 2002:120).
- Accomplish similar research for the PDRR and procurement phases for both RDT&E and procurement dollars (Sipple, 2002:120).
- Experiment with the sensitivity of the existing models by varying inputs (Sipple, 2002:120).
- Analyze database to explore and extract more predictor variables, to include higher-order terms and interactions, with potentially greater predictive capability.
- Allow time to pass under new Milestone structure and update the database to reflect the new structure and analyze the resulting effect on cost growth.

Appendix A – Logistic Regression Model

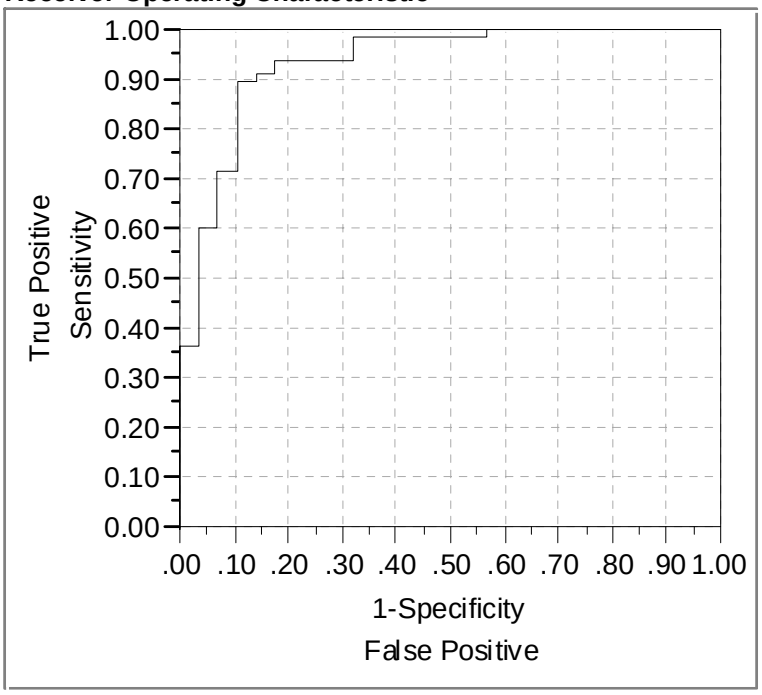
Nominal Logistic Fit for R&D (Total) Cost Growth?

RSquare (U)	0.5357
Observations (or Sum Wgts)	96

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-1.1278692	1.1774317	0.92	0.3381
Svs>3	3.80285403	1.5199046	6.26	0.0123
Maturity (Funding Yrs complete)	-0.4486886	0.1310089	11.73	0.0006
R&D Funding Yr Maturity % - squared	7.95441832	2.2275714	12.75	0.0004
Risk Mitigation?	-3.7384697	1.2431862	9.04	0.0026
EMD Prototype?	-2.001096	0.9319505	4.61	0.0318
Program have a MS I?	3.20026504	1.0356932	9.55	0.0020
Maturity (Funding Yrs complete)*(EMD Prototype?)	0.56873047	0.2117896	7.21	0.0072

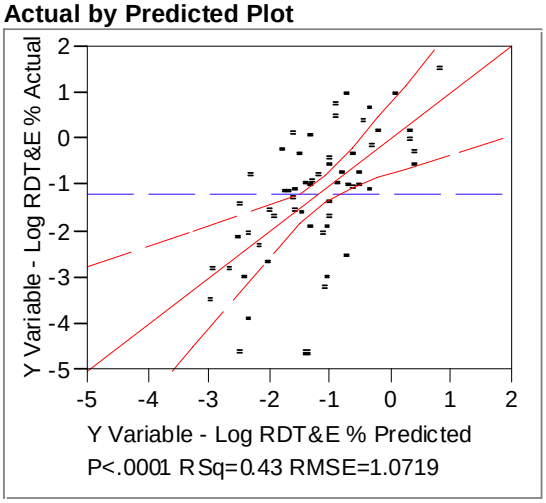
Receiver Operating Characteristic



Area Under Curve = 0.93435

Appendix B – Multiple Regression Model

Whole Model
Actual by Predicted Plot



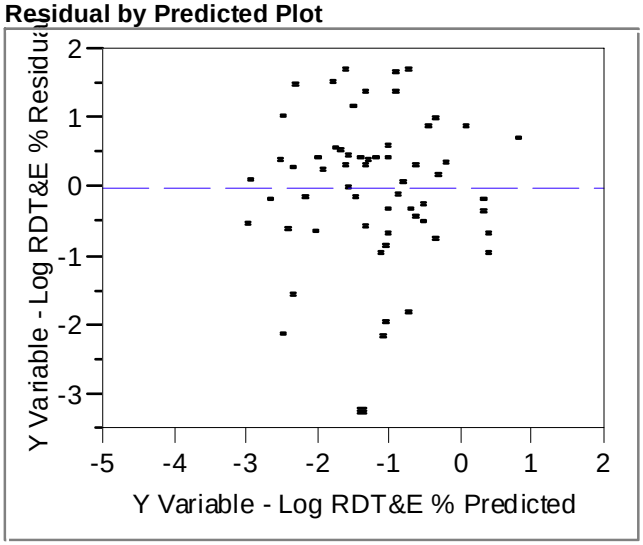
Summary of Fit

RSquare	0.426924
RSquare Adj	0.362047
Root Mean Square Error	1.071931
Mean of Response	-1.19824
Observations (or Sum Wgts)	60

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.070473	0.784431	-1.36	0.1781
Northrop Grumman	1.3557629	0.664538	2.04	0.0463
Funding Yrs of R&D Completed	0.132762	0.025576	5.19	<.0001
Maturity of EMD at IOC%	-1.929685	0.813505	-2.37	0.0214
Prototype?	0.8669499	0.346592	2.50	0.0155
Significant pre-EMD activity	-0.968515	0.325376	-2.98	0.0044
LRIP Planned?	0.7522629	0.302415	2.49	0.0160

Residual by Predicted Plot



Bibliography

Air Force Materiel Command. *AFMC Financial Management Handbook*. Wright-Patterson AFB OH: HQ AFMC, December 2001.

Bielecki, John V. *Estimating Engineering and Manufacturing Development Cost Risk Using Logistic and Multiple Regression*. MS thesis, AFIT/GCA/ENC/03-01. Graduate School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2003 (ADA-413231).

Calcutt, Harry M. *Cost Growth in DoD Major Programs: A Historical Perspective*. Executive Research Project. The Industrial College of the Armed Forces, National Defense University. Fort McNair Washington DC, 1993.

Coleman, R. L., J. R. Summerville, M. DuBois, and B. Myers. "Risk in Cost Estimating: General Introduction & the BMDO Approach." Briefing at the 33rd Annual DoD Cost Analysis Symposium. Williamsburg VA. 2 February 2000.

Drezner, J. A., J. M. Jarvaise, R. W. Hess, P. G. Hough, and D. Norton. *An Analysis of Weapon System Cost Growth*. Santa Monica CA: RAND, 1993 (MR-291-AF).

Jarvaise, J. M., J. A. Drezner, D. Norton. *The Defense System Cost Performance Database: Cost Growth Analysis Using Selected Acquisition Reports*. Santa Monica CA: RAND, 1996 (MR-625-OSD).

JMP® Version 5.0, (Academic), CD-ROM. Computer software. SAS Institute Inc., Cary NC, 2002.

Knoche, Chris. "Defense Acquisition Desk book: Selected Acquisition Report." Online Defense Acquisition Reference Guide. http://www.deskbook.osd.mil/default.asp. 1 July 2002.

Neter, John, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. Boston: McGraw-Hill, 1996.

Sipple, Vincent P. *Estimating Engineering Cost Risk Using Logistic and Multiple Regression*. MS thesis, AFIT/GAQ/ENC/02-02. Graduate School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 2002 (ADA-400576).

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
09-03-2004	**Master's Thesis**	Jun 2003 – Jan 2004

4. TITLE AND SUBTITLE

LOGISTIC AND MULTIPLE REGRESSION: THE TWO-STEP APPROACH TO ESTIMATING COST GROWTH

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Genest, Daniel C., First Lieutenant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GCA/ENC/04-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ASC/FMCE, AFMC
Attn: Mr. Michael J. Seibel
1865 4th St, Rm 134 DSN: 986-5478
WPAFB OH 45433-7123 e-mail: Michael.Seibel@wpafb.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study seeks to predict cost growth in major DoD acquisition programs using logistic and multiple regression. Specifically, this research uses logistic regression to determine whether or not cost growth will occur in a program and if so, then uses multiple regression to determine to what extent that cost growth will occur. We compile data from all defense departments using the Selected Acquisition Reports presented between 1990 and 2002. We combine the efforts of previous research and focus our study on cost growth in research and development dollars for the Engineering Manufacturing Development phase of acquisition. For the logistic regression portion of our research, we produce a seven-variable model that accurately predicts 72 percent of our randomly selected validation data. For multiple regression, we produce a six-variable model that accurately predicts the amount of cost growth incurred for 91 percent of those programs that do incur cost growth. We conclude that the two-step regression methodology offers a significant advantage over traditional methods by removing those data points that do not incur cost growth. We further conclude that there is no significant advantage gained by either isolating each cost variance category individually or by combining these categories.

15. SUBJECT TERMS
Logistic Regression, Multiple Regression, Cost Variance, Cost Growth, Selected Acquisition Report, SAR, DoD Cost Growth, Cost Growth in DoD Acquisition Programs, Predicting Cost Growth

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
REPORT	ABSTRACT	c. THIS PAGE			Edward D. White, PhD, (ENC)
U	U	U	UU	67	19b. TELEPHONE NUMBER *(Include area code)* (937) 785-3636, ext. 4540; e-mail: Edward.White@AFIT.edu

Standard Form 298 (Rev: 8-98)
Prescribed by ANSI Std. Z39-18